Nonvolatile Semiconductor Memory Technology

A Comprehensive Guide to Understanding and Using NVSM Devices

Edited by

William D. Brown
University of Arkansas

Joe E. Brewer
Northrop Grumman Corporation

IEEE
PRESS

IEEE Press Series on Microelectronic Systems
Stu Tewksbury, *Series Editor*
IEEE Solid-State Circuits Council, *Sponsor*
IEEE Components, Packaging, and Manufacturing Technology Society, *Sponsor*

The Institute of Electrical and Electronics Engineers, Inc., New York

This book and other books may be purchased at a discount
from the publisher when ordered in bulk quantities. Contact:

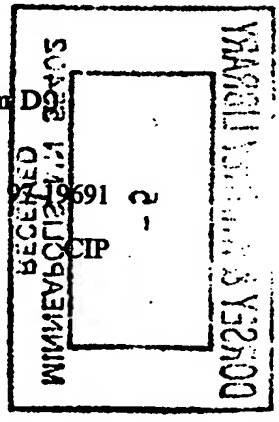
IEEE Press Marketing
Attn: Special Sales
445 Hoes Lane, P. O. Box 1331
Piscataway, NJ 08855-1331
Fax: (732) 981-9334

For more information on the IEEE Press,
visit the IEEE home page: http://www.ieee.org/

Printed in the United States of America
10 9 8 7 6 5 4 3

ISBN 0-7803-1173-6
IEEE Order Number: PC5644

Library of Congress Cataloging-in-Publication Data
Nonvolatile semiconductor memory technology : a comprehensive guide to understanding and using NVSM devices / edited by W. D. Brown, Joe E. Brewer.
p. cm. – (IEEE Press series on microelectronic systems)
Includes bibliographical references and index.
ISBN 0-7803-1173-6 (cloth)
1. Semiconductor storage devices. I. Brown, W. D. (William D., (date) . II. Brewer, Joe (Joe E.) III. Series.
TK7895.M4N634 1997
621.39'732–dc21
19691 2
CIP



content f any previously written cell. For all recent EEPROM circuits, the danger for soft-write does not exist because of the use of an isolating select transistor per memory cell.

Another special case of programming is the "read-disturb." During read-out of the information, voltages have to be applied to the nonvolatile memory transistor. These voltages can induce a threshold voltage shift in the cell that is addressed, as well as in some other cells. Read-disturb is mainly a concern for EPROM (and Flash EEPROM) memories for the same reason as noted above.

As an example, Fig. 1.48 shows the programming (both write and erase) characteristics of a FLOTOX-type nonvolatile memory transistor. For more details on the transient characteristics of the various cells, the reader is referred to the other chapters of this book.

1.6.2 Endurance Characteristics

With respect to overall reliability, two different features of the nonvolatile memory have to be considered. Nonvolatile memories can be reprogrammed frequently, but, in contrast to RAM memories, each write operation introduces some sort of permanent damage. This implies that the total number of write operations is limited; for example, most commercially available EEPROM products are guaranteed to withstand, at most, 10^4 programming cycles. The damaging of the memory cell during cycling is normally referred to as "degradation" and the number of cycles the memory can withstand is normally called its "endurance." Another failure mode of the nonvolatile memory is retention failure. This failure mode is discussed in the next section.

1.6.2.1 Floating Gate Devices. The program/erase endurance of floating gate devices is determined by four phenomena: tunnel oxide breakdown, gate oxide breakdown, trap-up, and degradation of the sense transistor characteristics. Whereas the first two are self-explanatory, trap-up is defined as the trapping of electrons in the oxide during programming operations. These trapped charges change the injection fields and thus, the amount of charge transferred to and from the floating gate during programming. This eventually leads to a situation where the difference in threshold voltage in the two possible memory states is so small that the sense circuit can no longer discern the two states. Degradation of the transistor characteristics occurs when CHE injection is used for programming. CHE injection is used primarily in EPROMs where endurance is restricted to a low number of cycles. Fowler–Nordheim injection also causes degradation of the transistor characteristics, but most devices make use of a separate tunnel area, leaving the sense transistor unaffected by programming operations.

As described by Mielke et al. [1.25], the main cause of endurance failure in TPFG devices is believed to be the trapping of electrons in the tunnel oxide (called trap-up), while thin oxide devices fail mainly because of thin oxide breakdown induced by very high oxide fields during programming. Trap-up also occurs in

2.2. MEMORY ARRAY CIRCUITRY

Incorporating EEPROM cells into large-scale memory arrays presents some unique problems because of the high voltages needed during programming operations. Although the overall memory organization of EEPROMs is similar to that of other volatile and nonvolatile memories—consisting of address/data buffers, row/column decoders, sense-amps, and other control circuits—some key issues have to be resolved before successful matrix implementation is possible—for instance, cell-disturb problems, high-voltage load circuits, over-erase protection circuits, programming timer circuits, and on-chip charge pumps with waveform shaping for 5 V-only operation. In the following discussion, V_{pp} refers to the programming voltage. Its exact value depends on the implementation technology.

When the individual cells (discussed in Section 2.1) are placed in a large memory matrix, the high electric field present during the programming operation can cause *disturb problems* in which programming of a selected cell affects the state of other unselected cells in the matrix. The simplest organization is shown in Fig. 2.14, where a memory transistor has been placed at each cross-point. To write the selected cell, V_{pp} is applied to the selected wordline and the column is grounded. Unselected wordlines are grounded, and columns are connected to V_{pp}, which will cause erasing of the unselected cell. Apart from the disturb problems, a large I_{pp} current will flow from the charge pump in unselected columns.

The disturb problem can be avoided if two transistors are used per memory bit—select and memory transistors, as shown in Fig. 2.15. To write the selected transistor, the selected wordline and column are connected to V_{pp} and ground, respectively, and the unselected wordlines are grounded and the columns pulled to V_{pp}. The select transistor disconnects the drain of the memory transistors of the

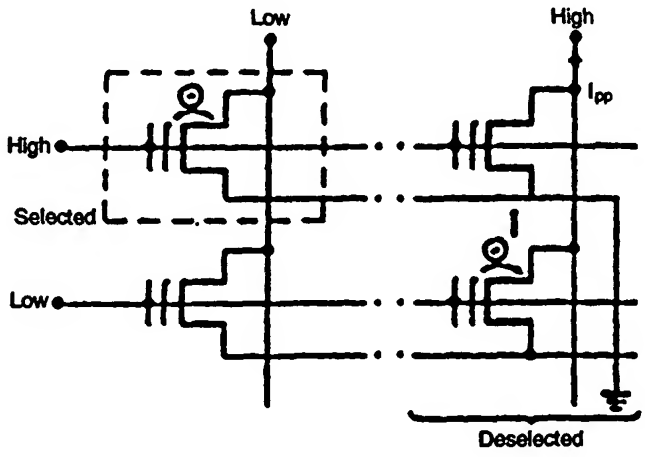


Figure 2.14 Disturb problems in single-transistor EEPROM arrays [2.16].

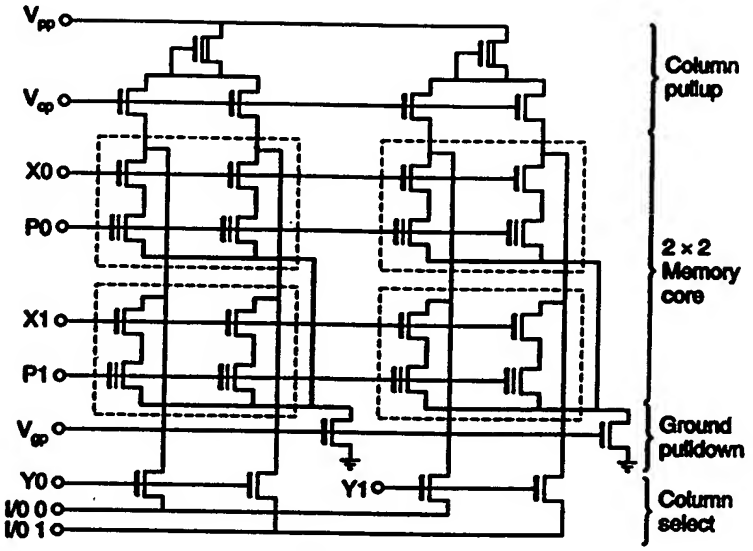
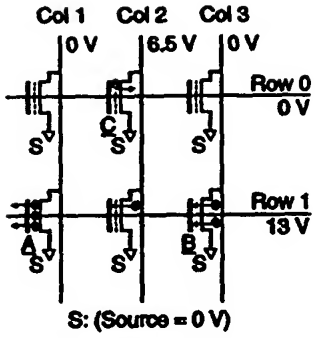


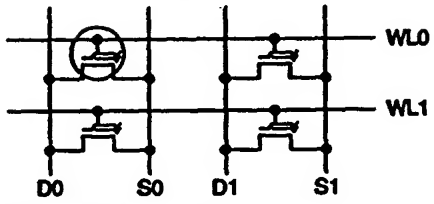
Figure 2.16 Four-byte EEPROM array organized in two rows, each constructed of two bytes (only two bits of the byte are shown) [2.16].

To minimize disturb problems encountered in the simple array given in Fig. 2.14, the half-voltage programming technique can be used, as shown in Fig. 2.17*a*. Here, half voltages, $V_{pp}/2$, are applied to unselected columns and rows, which results in only half the voltage necessary for tunneling being applied across the tunnel oxides of the unselected cells. Since the tunneling current depends strongly on the oxide field, the disturb problem is significantly reduced using this method. The three disturbs that can occur during programming are (a) DC erase, (b) DC program, (c) and program-disturb [2.17]. DC erase occurs on already written cells (i.e., with electrons in the floating gate) sharing the same wordline as the cell being written. During the write operation, the common wordline is high, and, if the electric field across the interpoly oxide is large enough, electron tunneling can occur, resulting in threshold voltage reduction. DC program occurs on cells in the erased state. Raising the wordline voltage of these cells creates a high field across the tunnel gate dielectric, which may cause electron tunneling to the floating gate. A written cell, sharing a column with another cell being written, will experience high electric fields between the floating gate and drain, which may cause electron tunneling from the floating gate to the drain. This effect, called program disturb, leads to reduced cell threshold voltage. This stress occurs at lower voltages (6.5 V) compared to those during DC program and DC erase.

A recent memory matrix implementation using this technique is shown in Fig. 2.17*b* [2.18, 2.19]. Applied voltages for write and erase are given in the figure. All cells are erased simultaneously by applying a negative high-voltage pulse (−11 V) to all wordlines and 5 V bitlines. By applying a negative voltage to the wordline during erase, instead of a high positive voltage to the source-drain junction, graded junc-



(a)



Cell (0,0) operating conditions

	WL0	WL1	S0	D0	S1	D1
Erase	−11 V	−11 V	5 V	Float	5 V	Float
Program	17 V	7 V	0 V	Float	7 V	Float
Read	3.3 V	0 V	0 V	1.5 V	0 V	1.5 V

(b)

Figure 2.17 (*a*) Schematic description f the disturb mechanism in single-transistor arrays using the half-voltage scheme: Cell A experiences DC erase disturb, Cell B experiences DC program-disturb, and Cell C experiences program-disturb [2.17], and (*b*) matrix implementation of the single-transistor memory cell [2.18].

tions are not needed. (A graded junction is required for a high breakdown voltage, but its drawback is that the cell area cannot be scaled easily.)

2.2.1 Charge-Pump Circuits

The recent development of 5 V-only memories requires that the programming voltage V_{pp} be generated using on-chip charge pumps. The basic voltage multiplier circuit is shown in Fig. 2.18 [2.20]. The nodes of the diode chain are coupled to the inputs via capacitors in parallel so that the capacitors have to withstand the full voltages developed across them. Efficient multiplication can be achieved with relatively high values of stray capacitance, and the current drive capability is independent of the number of multiplier stages. The multiplier operates by pumping charge packets along the diode chains as the coupling capacitors are successively charged and discharged each half-clock cycle. The voltages are not reset after each pumping cycle, so that the average node potentials increase progressively from the input to the output of the diode chain. The circuit implementation is shown in Fig. 2.19, where MOS transistors are used in the diode configuration. The clocks are generated by two inverters driven from an oscillator circuit. Typically, the utput is limited by MOS diodes to prevent output voltages from exceeding process limits, such as junc-

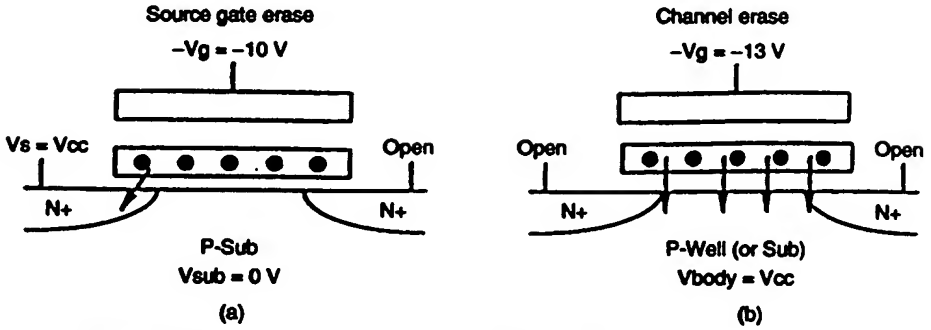


Figure 4.13 Cross-sectional drawing of (*a*) negative gate, floating gate to source erase, (*b*) negative gate, floating gate to channel erase [4.104].

4.2.3.2 Negative (Control) Gate, Floating Gate to Drain Erase. A triple-poly, split-gate Flash memory cell based on source-side injection for programming and negative control gate for erase (electrons tunnel from floating gate through the oxide to the drain) is shown in Fig. 4.14 [4.78]. The cell consists of a pair of stacked gates and a polycide blanket split gate. The select gate controls the weakly-on region during programming and prevents unselected cells from conduction in the program and read modes. The select gate runs along the channel direction. This eliminates the need for below-feature-size poly, reduces select gate delay, and improves the speed over select gate cells of the side-wall type [4.28, 4.88]. The use of a select gate allows the cell to operate in the depletion mode. To achieve a high read current in the erased state, the cell is put into the depletion mode.

The source-side reverse read is used to eliminate read-disturb induced by hot electrons. Figure 4.15 shows the array schematic and its operating conditions. To avoid high-voltage stress in the tunnel oxide during write, the high-voltage control gate (V_{pp}) and bitline (V_{cc}) are arranged to run parallel so that, in the program mode, both drain voltage and control gate voltage are present in the selected column. The self-imposed counterbiasing effect reduces the write-disturb.

The bitlines consist of source-lines and drain-lines arranged in alternating columns. An erase sector is naturally formed along each drain-line without any layout overhead. Each sector consists of two columns of opposing cells that share a common control gate.

	Vd	Vcg	Vsg	Vs	Vt/Ir	Time
Write	5V	13V	1.8V	0V	>5V	10 μS
Erase	5V	−12V	0V	F	−2V	1 sec
Read	0V	3.8V	5V	2	130 μA	

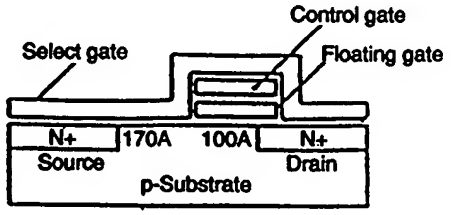


Figure 4.14 Cell cross-sectional view and operating conditions [4.78].

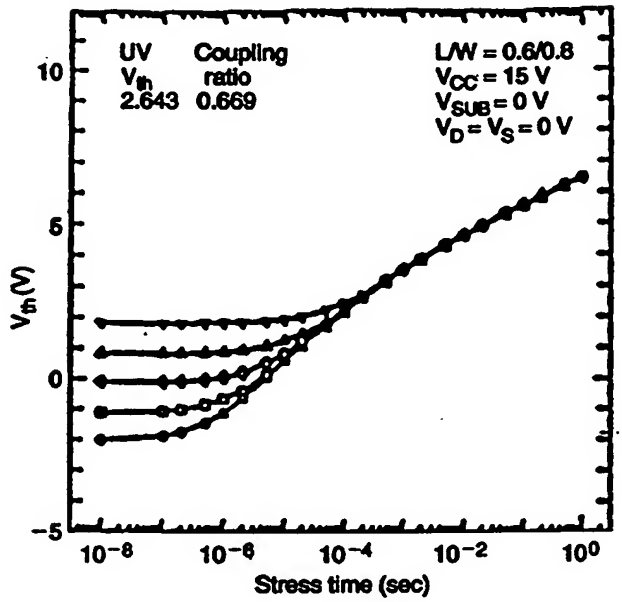


Figure 4.28 F–N programming (electrons emitted from the Si/SiO_2 interface) characteristics of a cell. F–N programming characteristics have good convergence of the programmed V_{tm} regardless of the initial V_{tm} [4.53].

4.2.6 Program-Disturb Mechanisms

The two principal memory cell-disturb mechanisms that can occur during programming of an array are called gate-disturb (DC program) and drain-disturb (program disturb) [4.21]. These mechanisms can occur in memory cells sharing a common wordline (WL) or a common bitline (BL) while one of the cells is being programmed. The effect of these disturbs on the different cells of the memory array is shown in Fig. 4.32.

Gate-disturb occurs in unprogrammed or erased cells that are connected to the same wordline as the cell being programmed. These cells have a low cell threshold voltage. During the programming operation, the common wordline is connected to a high voltage. The electric field across the tunnel oxide becomes high and may cause tunneling of electrons to the floating gate from the substrate. The threshold voltage of the cell will increase and reduce the sense margin. In severe cases, the cell is programmed unintentionally.

Drain-disturb occurs in programmed cells that are on the same bitline as the cell being programmed. These cells will experience a high electric field between the floating gate and the drain. This may cause electrons to tunnel from the floating gate to the drain and lead to a reduced cell threshold voltage.

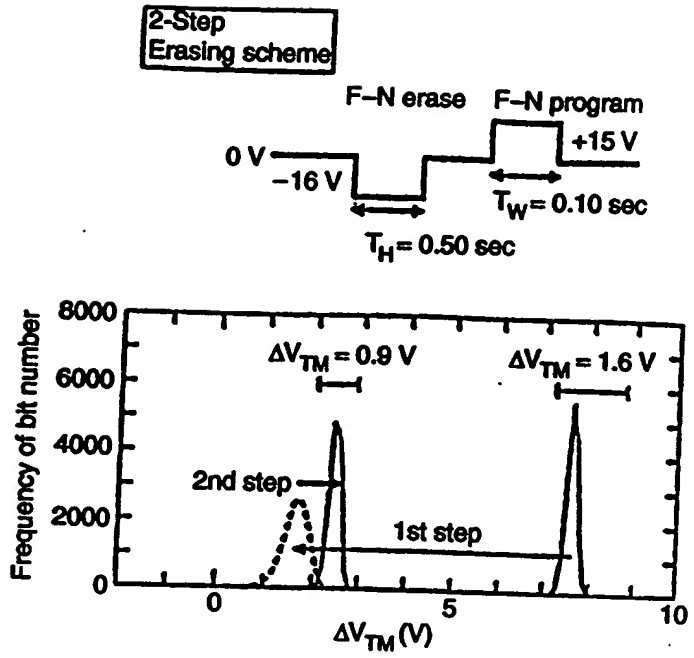


Figure 4.29 The erased-V_{tm} distribution of 16 Kbit cell array using two-step erasing. The erased V_{tm} distribution after two-step erasing operation is suppressed from ΔV_{tm} = 2.0 V to 0.9 V in width [4.53].

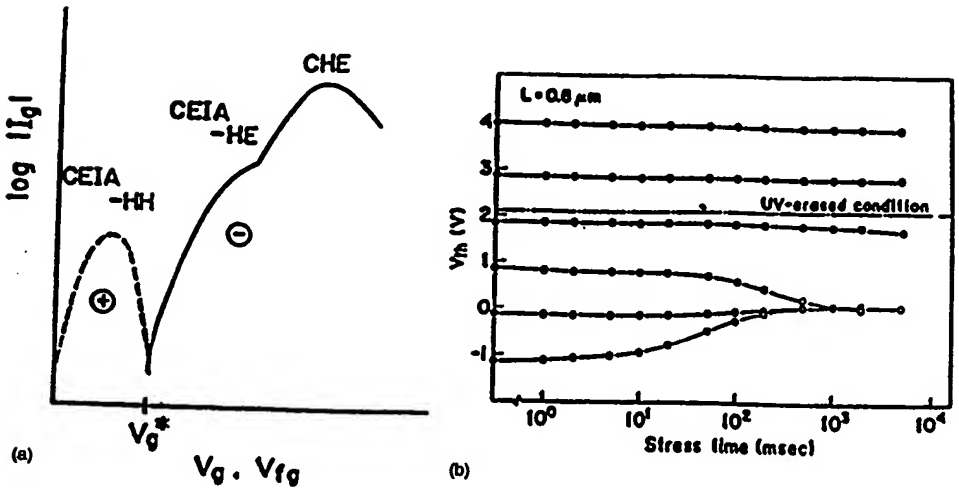


Figure 4.30 (*a*) Characteristics of gate current, I_g, versus gate voltage, V_g, for a NMOSFET, (*b*) threshold voltage, V_{th}, versus drain-stress time with different starting threshold voltages as parameters [4.50].

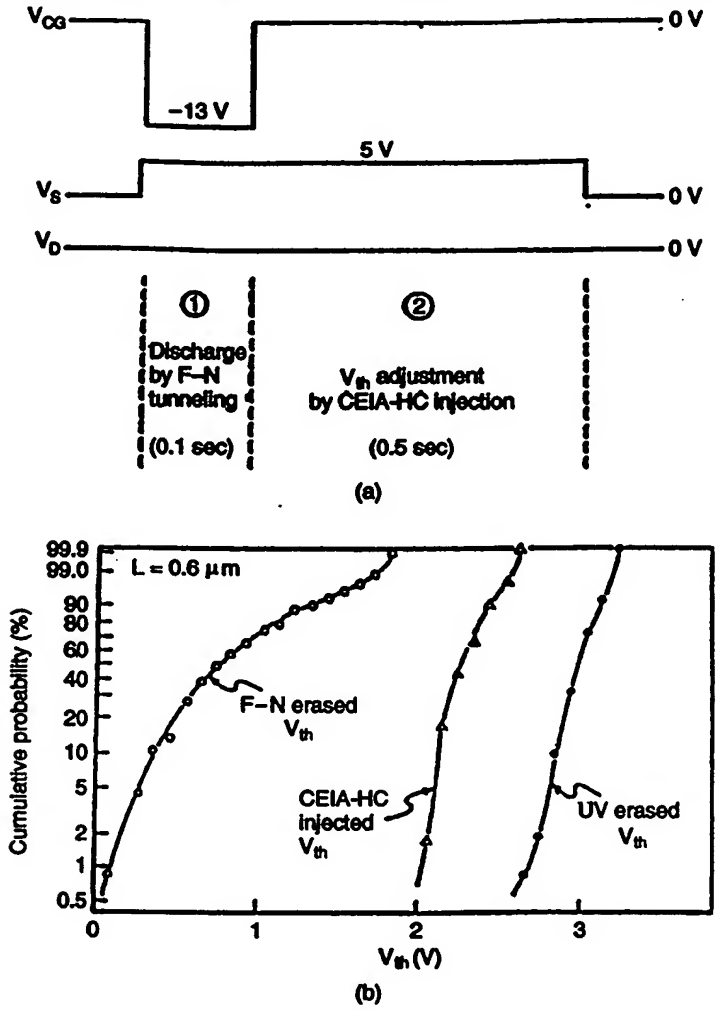


Figure 4.31 (*a*) Timing diagram used in experiments on new erase sequence, (*b*) threshold voltage distribution before and after adjustment by CEIA-HC injection [4.50].

An important design consideration is the proper selection of the programming voltages to minimize these disturbs. Write/erase cycling of the cell also affects these disturbs. Verma and Mielke [4.21] showed that the drain-disturb characteristics of Flash memory devices are excellent with no measurable change until several thousand cycles. However, write/erase cycling has an influence on the gate-disturb behavior. Figure 4.33 shows the gate-disturb time as a function of cycling. Before cycling, the disturb time is about 100 seconds. After 100 cycles, this margin is decreased about two to three rders f magnitude. This degradation in gate-disturb sensitivity is caused by hole trapping during erase [4.21].

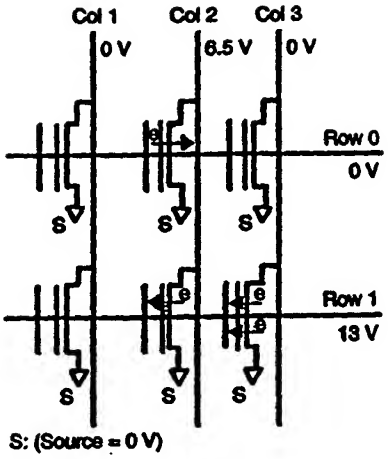


Figure 4.32 Schematic description of disturb during programming [4.21].

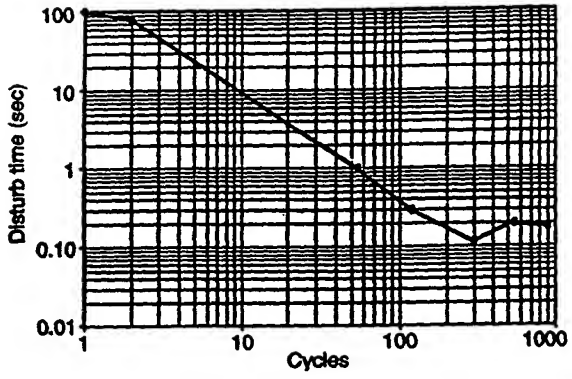


Figure 4.33 Gate-disturb time as a function of cycling [4.21].

4.3. FLASH MEMORIES WITH CHANNEL HOT-ELECTRON PROGRAM AND POLY-TO-POLY ERASE

Even though erasing through the source or drain does not require an extra transistor terminal, thereby giving the most compact cell, the use of junctions does create the problem of breakdown current. Also, a thin oxide is required for the cell gate oxide. Instead of erasing through the silicon, it is also possible to have an extra polysilicon

electrode that connects through an oxide to the floating gate. This extra poly electrode results in a slightly larger memory cell or a more complicated process.

The poly-to-poly erase (floating gate to control gate in a double-poly structure, or floating gate to an erase gate in a triple-poly structure) avoids stressing of the gate oxide during erase. The first modern Flash EEPROM [4.3–4.5] from Toshiba used poly-to-poly erase and will be described first, followed by a contactless, virtual ground array version from Sandisk. In both cases, erase occurs from the floating gate to an erase gate, although the locations of the erase gate in the cell are different.

4.3.1 Triple-Poly NOR

Triple polysilicon Flash memory cells [4.3–4.5] are shown in Figs. 4.34, 4.35, and 4.36. The select gate is integrated with the floating gate (merged pass gate). The first polysilicon layer is used as an erase gate and is located between the field oxide and the floating gate. The second polysilicon layer is used as a floating gate in a manner similar to an UV EPROM. The third polysilicon layer is used as a control gate and is, in fact, used as a bit select-line for programming and reading.

The cell is programmed by a channel hot-carrier injection mechanism similar to EPROM. The erasure is accomplished using the first level of polysilicon which serves as an erase electrode causing field emission of electrons from the bottom of the floating gate. In the triple-poly NOR cell, as shown in Fig. 4.34, one part of the channel is controlled by the control gate. The third polysilicon layer is used both as a gate of the selection transistor and a control gate of the triple-poly NOR cell. This selection transistor enables the triple-poly cell to remain in the enhancement mode even if the floating gate transistor becomes a depletion mode device after erasure. The erase gate of the memory cell is supplied with the boosted voltage (V_{EG}), which enables field emission from the floating gate. This boosted voltage (V_{EG}) is produced from the program voltage (V_{PP}). All of the memory cells are erased simultaneously by the erase gate because they are commonly connected to each other.

The read data depend on the stored electric charges on the floating gate similar to the case for EPROM. The erased memory cell (the floating gate not charged with electrons) conducts by applying voltage as shown in Fig. 4.36. The virgin memory cell, prior to being programmed, has the same threshold voltage as the erased memory cell. However, their g_m values are different. The threshold voltage of the F-E2PROM cell is controlled by the selection transistor. However, the g_m depends on the value of the stored electric charges in the floating gate, even if the memory cell is erased.

4.3.2 Triple-Poly, Virtual Ground Contactless [4.63]

The Flash memory cell uses a triple-polysilicon, single-metal, split-channel structure with buried n^+ diffused source-drain, integrated into a contactless, virtual ground array architecture. The buried diffusion bitlines are contacted periodically

TABLE 4.5 MEMORY CELL OPERATION CONDITIONS

Operation	Wordline	Source-Line	Bitline
Erase	14 V	0 V(V_{ss})	0 V(V_{ss})
Program	V_t	11 V	0 V data "0" V_{cc} data "1"
Read	V_{cc}	0 V(V_{ss})	2 V

4.3.3.2 Erasing. The cell is erased by Fowler–Nordheim tunneling of electrons from floating gate to control gate through interpoly oxide. During erase, the source and drain are grounded and the wordline is raised to 14 V. The conditions for erase are given in Table 4.5. The low coupling ratio between the control gate and the floating gate provides a significant voltage drop across the interpoly oxide, which is the same everywhere between poly1 and poly2. A high field is generated primarily in the area of the tunneling injector. Charge transfer is very rapid and is eventually limited by the accumulation of positive charge on the floating gate. This positive charge raises the floating gate voltage such that there is insufficient voltage drop across the poly-to-poly dielectric to sustain Fowler–Nordheim tunneling.

The removal of charge can leave a net positive charge on the floating gate. The positive charge on the floating gate reduces the memory cell's threshold voltage to about the select gate V_t. The applied sense voltage is sufficient to turn on both the select transistor and the memory transistor in the addressed memory cell.

Erase can either be by fixed program pulses generated by an internal timer or algorithmically generated by an external controller to optimize erase conditions. Internal verify circuits assure an adequate erase margin.

4.3.3.3 Erase Disturb. Enhanced-field tunneling injector devices are internally organized by pairs (pages) of even and odd rows. Each row pair (page) shares a common source-line and has the wordline at the same voltage potential during erase. Thus, all bytes along the common wordlines are erased simultaneously. All other wordlines (pages) do not receive the erasing high voltage. Therefore, erase-disturb is not possible. The column leakage phenomenon caused by "overerase" in 1-T cells is avoided because the split gate provides an integral select gate to isolate each memory cell from the bitline.

4.3.3.4 Programming. The cell is programmmed using high-efficiency source-side channel hot-electron injection. The conditions for programming are given in Table 4.5. During programming, a voltage of a cell threshold of approximately V_T volts is placed on the control gate via the wordline. This is sufficient to turn on the channel under the select portion of the control gate. The drain is at

approximately V_{ss} if the cell is to be programmed. If the drain is at V_{cc}, programming is inhibited. The drain voltage is transferred across the select channel because of the voltage on the control gate. The source is at approximately 12 V. The source to drain voltage differential (i.e., 11 V—approximately V_{ss}) generates channel hot electrons. The source voltage is capacitively coupled to the floating gate. The electric field between the floating gate and the channel sweeps the channel hot electrons that cross the Si–SiO_2 barrier height of approximately 3.2 eV to the floating gate very efficiently.

The programming effect is eventually self-limiting as negative charge accumulates on the floating gate. The programming source-drain current is low. Thus, the source voltage can be generated by an on-chip charge pump. The program time is fast because of the relatively high efficiency of source-side injection. Programming can be by fixed program pulses generated either by an internal timer or by an external controller to optimize program conditions.

4.3.3.5 Program-Disturb. There are two possible types of program-disturb with the field-enhanced tunneling injection cell: reverse-tunnel disturb and punch-through disturb.

Reverse-tunnel disturb can occur for unselected erased cells sharing a common source-line, but on the other row of the selected page to be programmed. Thus, the wordline is grounded. The source voltage is capacitively coupled to the floating gate of the unselected erased cell. If there is a defect in the oxide between the control gate and the floating gate, Fowler–Nordheim tunneling may occur. This could program the unselected erased cell.

Punch-through disturb can occur for selected erased cells, that is, those sharing a common source-line and wordline in an adjacent inhibited bitline. An inhibited bitline is taken high to prevent normal channel hot-electron injection. If there is a defect that reduces the bitline voltage or creates punch-through along the select gate channel, hot electrons could be available to program the inhibited erased cell.

Proper design and processing can prevent both mechanisms. Devices with this memory architecture do not have program-disturb caused by accumulated erase/programming cycles because each page is individually isolated. Each cell is only exposed to high voltage within the selected page along the row- or source-line. There is no high voltage on the bitline.

4.4. FLASH MEMORIES WITH FOWLER–NORDHEIM TUNNEL PROGRAM AND ERASE

Although EPROM-like cell structures offer density and historical learning advantages for Flash memory implementation, cell structures based on Fowler–Nordheim tunneling offer low power and single, low-voltage supply advantages. Examples are DINOR [4.58], AND [4.57], NAND [4.14], and HiCR [4.68]. Based on current understanding, the single-voltage supply Flash memories, below 3 V V_{cc}, will require

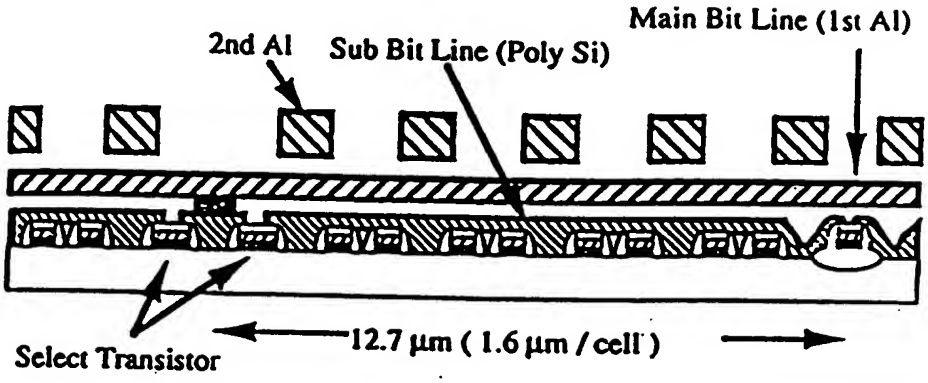


Figure 4.51 Schematic drawing of the DINOR cell [4.58].

in terms of wafer processing simplicity. The first aluminum layer forms the main bitline. The second aluminum layer in the cell array is used to stitch the gates of the select transistors and the main wordlines.

The sector size of the DINOR is 1 KB in which the cells are arrayed in 16 bits (=2 basic units described above) × 512 bits (connected to the same wordline).

4.4.3.3 Disturb and Endurance Characteristics. The DINOR cell solves disturb problems by utilizing the divided bitline and parallel programming. By setting all select transistors off except for the one that is in the programming mode, the maximum drain-disturb is only seven write cycles from the cells connected to the same sub-bitline. The maximum gate-disturb is only one write cycle due to the use of a latch circuit. Characteristics of drain-disturb and gate-disturb, with sufficient margin for device operation, have been demonstrated [4.58].

In the case of the DINOR cell, one more new disturb mode, substrate disturb, exists. Substrate disturb is the undesired erasure caused by V_{sub} (substrate voltage) stress during the erasure of other sectors. The longest time over which substrate-disturb occurs is of the order of one year for actual 16 Mbit devices. By applying an erase inhibit voltage (1/2 of V_{sub}) to the source-lines of the unselected sectors, the effective V_{sub} can be reduced by one-half, making the substrate-disturb immunity longer than 1E10 seconds.

4.4.3.4 Virtual Ground DINOR. A virtual ground DINOR array, utilizing F–N tunneling (for program and erase) has been described by using the asymmetrical offset source-drain structure [4.64] shown in Figs. 4.52 and 4.53. One side of the drain region has an offset against the floating gate, and the other side is overlapped with it. During programming, F–N tunneling current only flows to the overlapped drain region that is selected. Thus, only the selected cell is programmed. The operating conditions for the array are shown in Table 4.8.

During programming, the unselected bitline is kept open, and the unselected wordline is at 5 V to minimize program-disturb. During erase, 12 V is applied to the wordlines to be erased, and bitlines and source-lines are grounded. Because of the F–N tunnel program/erase operation, high voltage can be generated on-chip from the supply voltage. The memory cell array (Fig. 4.57) has the NOR structure. The select transistors act as switches for connecting and disconnecting the target block to a main data-line and a main source-line.

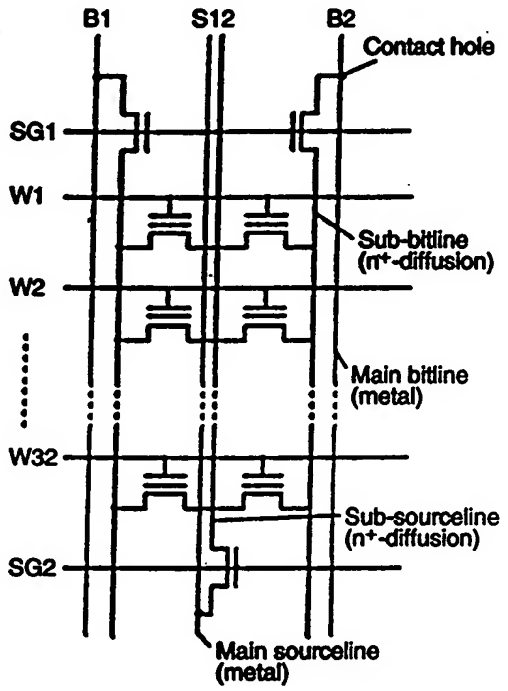


Figure 4.57 Schematic of HiCR memory array [4.68].

4.4.4.1 Program Disturb. Two principal disturbances can occur during programming of an array (Figs. 4.58 and 4.59). The first (I) is unexpected programming of the cell (12) connected to the same source-line (S) and to the selected wordline (W1). It is caused by the source voltage, $V_{CG} - V_{TM}$, of the cell, where V_{CG} is the program inhibit voltage of the unselected wordline (W2) and V_{TM} is the threshold voltage of the erased cells (cell 21 works as a pass transistor when it is erased). The second disturbance (II) is the unexpected programming of cells (21, 22, . . .) connected to the selected bitline (B1).

These problems can be solved by utilizing parallel programming and sub-bitlines. Disturbance I, which has a maximum duration of only one write cycle (1 ms), can be solved by the use of a latch circuit [4.71]. The solution to disturbance II,

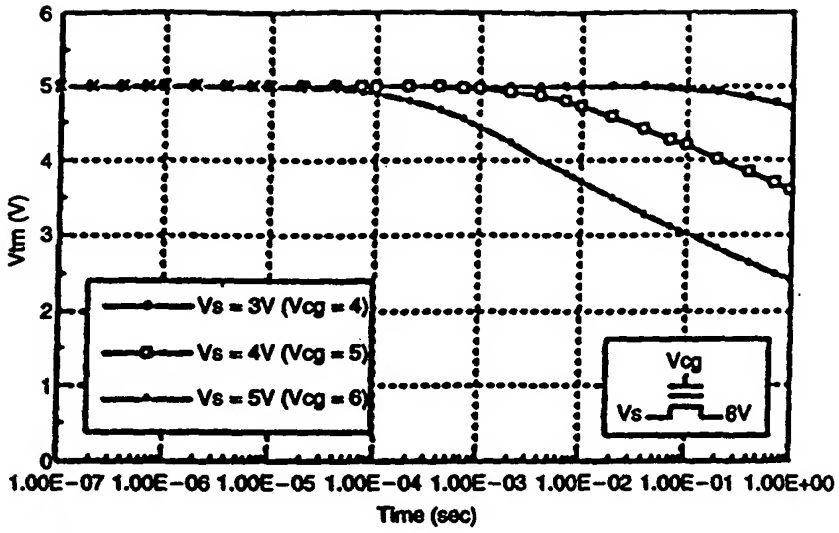


Figure 4.58 Threshold-voltage shifts during disturbance: I. The cell was programmed to a V_t of 5 V before disturbance measurement. Maximum disturbance I is 1 write cycle (1 ms) [4.68].

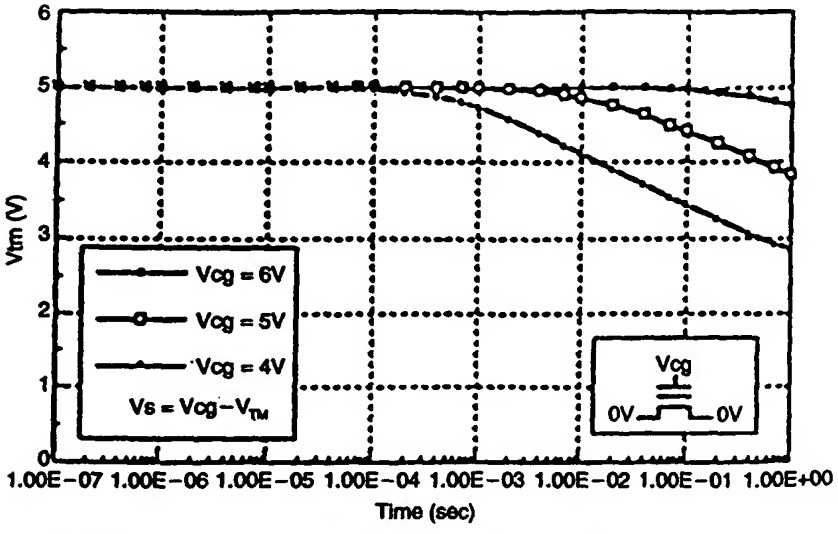


Figure 4.59 Threshold-voltage shifts during disturbance: II. Maximum disturbance II is 31 write cycles (31 ms) [4.68].

which has a maximum durati n of 31 write cycles (31 ms), can be realized by setting the select transistors off except in the programming mode. Characteristics f disturbances I and II, which have sufficient margin for device operation, are shown in Figs. 4.58 and 4.59, respectively.

4.4.5 NAND

In the drive to reduce cell size, eliminating the select transistor in a full-feature EEPROM results in a bulk/block erasable Flash EEPROM. The Flash EEPROMs discussed so far are the NOR structure, in which memory cells are connected to a bitline in a parallel manner, and the NAND structure, which reduces the cell size by connecting the cells in series between a bitline and a source-line, thus eliminating the contact hole [4.13, 4.26, 4.32]. The resulting cell structure occupies 85% of the NOR cell area of a stacked gate array.

4.4.5.1 NAND Structure. Figure 4.60 shows the layout and the equivalent circuit of the NAND-structured cell. As shown in the figure, the NAND-structured cell arranges eight or sixteen memory transistors in series, sandwiched between two select gates, select gate 1 (SG1) and select gate 2 (SG2). The first gate (SG1) ensures selectivity, and the second (SG2) prevents the cell current from passing during a programming operation. The floating gates are made of first-level polysilicon. The control gates, which are wordlines in an array, are made of second-level polysilicon. The dielectric between the floating gate and the control gate is an ONO stack.

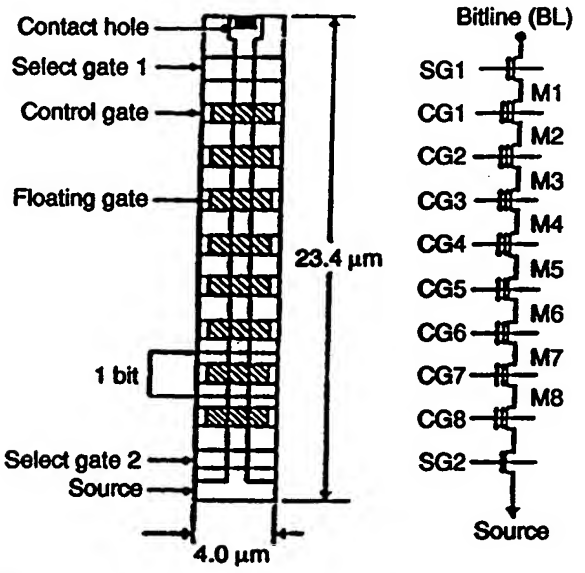


Figure 4.60 Top view and equivalent circuit of NAND-structural cell [4.13].

ZERO data programming cell is grounded. Electrons are injected from p-well(2) to the floating gate by F–N tunneling. The V_t of the selected cell becomes positive. Ten volts is applied to the bitline of the ONE data programming cell. The V_t of the unselected cell remains negative because the voltage across the tunnel region is inadequate to start the tunneling current. In this case, while 10 V is applied between the drain and p-well(2), gated-diode breakdown does not occur because the control gate is biased at 10 V. In both erase and program operation, the gated-diode breakdown is suppressed entirely. As a result, all high voltages are internally generated easily from a 5 V power supply through charge-pump circuits because there is no breakdown leakage current. Furthermore, oxide degradation [4.9] from hole trapping, due to gated-diode breakdown, is avoided. This results in improved write/erase endurance.

4.4.5.3 Disturb Mechanisms. The operating conditions, including the inhibit conditions, are shown in Fig. 4.63. Erasure of deselected cells in the deselected NAND block is prevented by applying a high voltage of the order of 20 V to the wordlines. Thus, the tunnel dielectric field is zero in the deselected NAND block, and no erase-disturb condition exists.

In order to prevent programming of deselected cells sharing the same bitline in the same NAND block, the control gate voltage is raised to V_m (a medium voltage) to reduce the tunnel dielectric field between the floating gate and the channel. Programming of deselected cells sharing the same control gate is inhibited by applying a medium voltage (V_m) of around 7 V to the deselected bitline to reduce the electric field between the floating gate and the channel.

Programming of cells in each NAND block is performed in a serial order from the source-line side to the bitline side to avoid unintentional charging of deselected cells sharing the same control gate. This ordering prevents the existence of programmed cells on the drain side and avoids the channel potential decrease below 8 V of the deselected bitline voltage to minimize the disturb for deselected cells. Disturbs for deselected erased cells sharing the bitline exist, although the programming (disturb) time is fast enough to prevent the charge gain.

4.4.5.4 Special Features

Advantages

1. Since the number of bitline contacts in the NAND array is reduced by one-eighth to one-sixteenth compared to that of the standard T-cell Flash array, the unit cell size is smaller.
2. A NAND cell with n^+ source and drain junctions is more scalable than the E-Tox cell since the E-Tox cell has a graded-source junction.
3. Programming and erasing are achieved by F–N tunneling and need less power (i.e., low currents), thus allowing a single power supply peration by utilizing internal charge-pump circuits.

the drastically reduced current requirement during erasure, an optional on-chip charge pump can be built t enhance the erase speed to less than 1 msec.

4.5.1.2 Inhibit Conditions. A small section of the memory array during programming is depicted in Fig. 4.65. The selected bit to be programmed is encircled. The bits that share the same wordline will see the wordline voltage repeatedly during programming. The worst case bit will be subjected to this wordline voltage for a duration that equals the total time to program the other bits on the same wordline. The resultant threshold voltage gain is called the gate-disturb. The bits that share the same bitline will see the bitline voltage repeatedly during programming. The worst case bit will be subjected to this bitline voltage for a duration that equals the total time to program the other bits on the same bitline. The resultant threshold voltage loss is called the drain-disturb. As in most Flash EEPROM cells, the tunnel dielectric thickness in the floating gate channel region is around 100 Å. At this thickness and without a select transistor between the drain and the floating gate transistor, suppression of program-disturbs is achieved through the proper balance of the programming voltages and the programming speed. Typically, the duration of a particular half-selected memory bit, exposed to the programming bias conditions, is kept below 1 second.

4.5.1.3 Read. Reading of the SCSG cell is accomplished by raising the drain to about 1 V, grounding the source, and raising the control gate to the supply voltage. The charge state is determined by the relative magnitude of the cell current to a reference current in the sense amplifier. The split gate configuration allows the

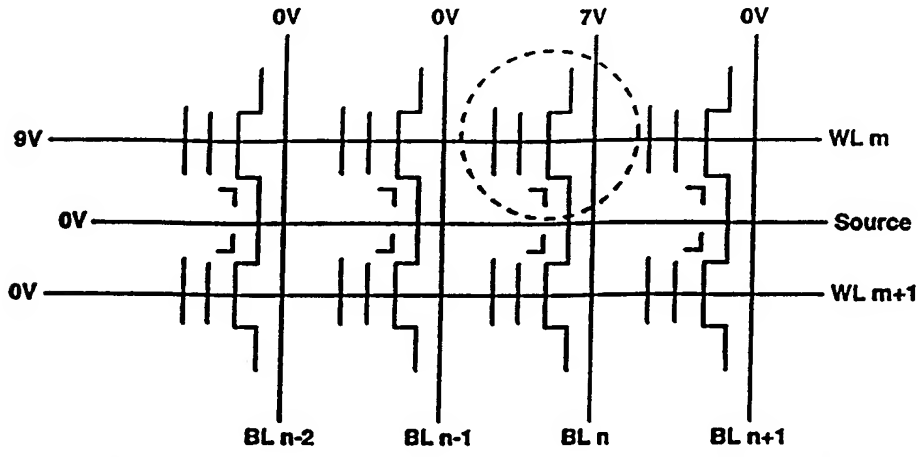


Figure 4.65 Program inhibit conditions [4.99]. Circled cell indicates the selected cell (row#m, column#n). The half-selected bits along the same wordline (WL m) will see the gate disturb, while the half-selected bits along the bitline (BL n) will see the drain disturb. Other bits are unaffected [4.158].

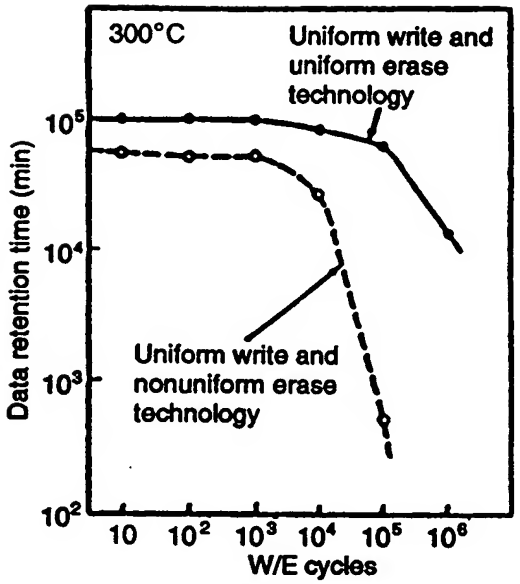


Figure 4.103 NAND data retention time [4.14].

4.6.9 Read-Disturb

The read-disturb stress condition occurs on erased bits that share a wordline with a bit that is being read. The common wordline places the control gate of the erased cells at 5.0 V (V_{cc}). The selected device's drain is driven to about 1.0 V. The unselected bits have their source, drain, and substrate at 0.0 V. In high-density Flash memories, many bits are put into a low-field stress condition when one bit is read. This condition is shown in Fig. 4.104 [4.135].

It has been reported that thin gate oxides exposed to high-field stress and high levels of charge injection can develop a pre-breakdown leakage condition [4.136, 4.137]. As program/erase cycling requirements increase, and the Flash tunnel oxide thickness is decreased, the EEPROM cell becomes more susceptible to tunnel oxide leakage [4.138]. During P/E cycling, some Flash bits can exhibit a leakage condition that permits charge gain on erased bits due to the low electric field present during a read operation. Under prolonged or DC read conditions, the defective cells can appear programmed.

It was shown that the charge gain takes place by electron tunneling through a corrupted oxide barrier [4.135]. The barrier reduction is caused by positive charge trapping at the tunnel oxide to source junction. The charge trapping is due to hole generation during F–N erase. Since the effective barrier takes on various levels, it was proposed that the configuration of trapped charge determines the extent of barrier lowering. The effective barrier of leaking cells was determined by tracking the cell threshold voltage during read-disturb stress conditions.

source, will have the greatest effect on barrier reduction [4.135], as sketched in Fig. 4.106, by reducing the tunneling distance into the conduction band of the floating gate. It is speculated that some trap configurations are more likely to occur, therefore favoring the appearance of the corresponding barrier height. The effective barrier ranged from 0.39 to 0.88 eV [4.135]. Figure 4.106 shows the conduction band of source, tunnel oxide, and floating gate regions under a read-disturb stress.

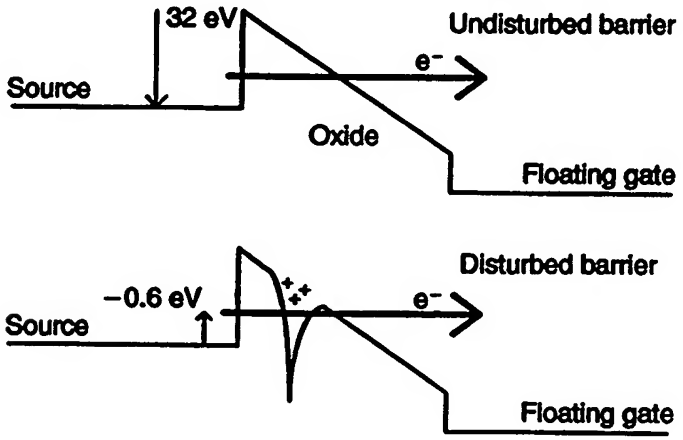


Figure 4.106 Sketch of the conduction band of the Flash source, tunnel oxide, and floating gate region under read-disturb stress. The disturbed barrier has a reduced tunneling distance due to trapped positive charges located near the source region [4.135].

4.7. PROCESS TECHNOLOGY

The process technology required for floating gate devices is derived from standard polysilicon gate technology that has been used so successfully in the manufacturing of SRAMs, DRAMs, and logic devices. The most significant addition is floating gate technology, which requires the highest quality insulating dielectric. Another important consideration is the high voltage that is required for EPROM programming and tunnel erase. It puts special requirements on isolation, junction breakdown, and transistor technology. A further reliability consideration is the quality of the dielectric used in the transistors and in the floating gates.

4.7.1 Floating Gate Technology

In a typical EPROM technology, poly1 is used exclusively for the floating gate. After poly1 is defined, a poly-to-poly dielectric is formed that surrounds the poly1